



(12) 发明专利申请

(10) 申请公布号 CN 118871918 A

(43) 申请公布日 2024. 10. 29

(21) 申请号 202380027529.X

(22) 申请日 2023.05.19

(30) 优先权数据

63/343,819 2022.05.19 US

(85) PCT国际申请进入国家阶段日

2024.09.13

(86) PCT国际申请的申请数据

PCT/EP2023/063488 2023.05.19

(87) PCT国际申请的公布数据

W02023/222884 EN 2023.11.23

(71) 申请人 渊慧科技有限公司

地址 英国伦敦

(72) 发明人 N·C·拉比诺维茨 N·A·罗伊

J·金

(74) 专利代理机构 北京市柳沈律师事务所

11105

专利代理师 金玉洁

(51) Int.Cl.

G06N 3/008 (2006.01)

G06N 3/044 (2006.01)

G06N 3/045 (2006.01)

G06N 3/084 (2006.01)

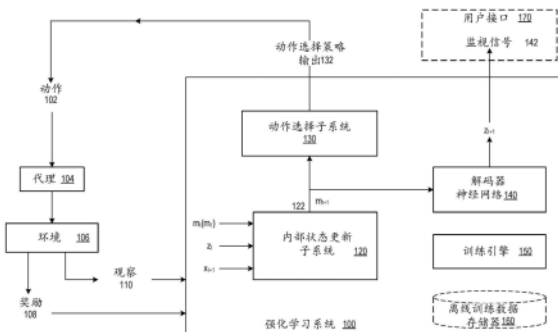
权利要求书5页 说明书16页 附图8页

(54) 发明名称

具有反事实干预的机器学习系统

(57) 摘要

用于训练和使用机器学习系统来控制代理执行任务的系统、方法和计算机程序。机器学习系统使用反事实内部状态来训练,使得它可以提供以因果术语解释系统行为的输出,例如,就其环境的使系统为代理选择特定动作的方面而言。



1. 一种训练机器学习系统的计算机实现的方法,所述机器学习系统被配置为监视环境中执行任务的代理的控制,其中,所述机器学习系统被配置为,对于一系列时间步骤中的每个时间步骤:

获得表征当前时间步骤的环境状态的观察;

使用所述机器学习系统的动作选择子系统来处理所述机器学习系统在所述当前时间步骤处的内部状态以生成动作选择策略输出,其中所述机器学习系统在所述当前时间步骤处的所述内部状态取决于所述当前时间步骤的所述观察;

使用解码器神经网络处理所述机器学习系统在所述当前时间步骤处的所述内部状态,以生成所述当前时间步骤的解码器输出,其中,所述当前时间步骤的所述解码器输出描述所述机器学习系统在所述当前时间步骤处的所述内部状态;

响应于所述观察,使用所述动作选择策略输出来选择要由所述代理在所述当前时间步骤处执行的动作;

使所述代理执行所选择的动作;以及

提供从所述当前时间步骤的所述解码器输出导出的信号作为用于监视所述环境中的所述代理的控制的监视信号;

所述方法包括通过以下方式训练所述机器学习系统:

通过使用内部状态更新子系统处理所述当前时间步骤的解码器输出、下一时间步骤的观察和反事实内部状态的组合来确定所述机器学习系统在下一时间步骤的内部状态,其中所述反事实内部状态不同于所述机器学习系统在所述当前时间步骤处的内部状态。

2. 根据权利要求1所述的方法,其中,训练所述机器学习系统还包括,对于所述一系列时间步骤中的每个时间步骤:

获得表征所述当前时间步骤的环境状态的观察;

使用所述机器学习系统的所述动作选择子系统来处理所述机器学习系统在所述当前时间步骤处的内部状态,以生成所述动作选择策略输出;

使用所述解码器神经网络处理所述机器学习系统在所述当前时间步骤处的内部状态,以生成所述当前时间步骤的解码器输出;

响应于所述观察,使用所述动作选择策略输出来选择要由所述代理在所述当前时间步骤处执行的所述动作;

使所述代理执行所选择的动作,其中,响应于使所述代理执行所选择的动作,所述机器学习系统接收奖励;以及

通过将强化学习目标函数的梯度反向传播到所述机器学习系统中,使用所述观察、使用动作选择策略输出选择的动作和所述奖励来训练所述机器学习系统;并且其中,所述训练包括:

通过使用所述内部状态更新子系统处理所述当前时间步骤的解码器输出、所述下一时间步骤的观察和所述反事实内部状态的组合来确定所述机器学习系统在所述下一时间步骤处的内部状态。

3. 根据权利要求2所述的方法,其中,通过处理所述当前时间步骤的所述解码器输出、所述下一时间步骤的所述观察和所述反事实内部状态的组合来确定所述机器学习系统在所述下一时间步骤处的所述内部状态定义了根据干预时间表执行的干预,其中,所述干预

时间表包括随机选择用于所述干预的时间步骤。

4. 根据权利要求2或3所述的方法, 其中, 所述训练包括: 对于第一时间步骤之后的多个干预时间步骤中的每一个, 通过使用所述内部状态更新子系统处理当前时间步骤的解码器输出、下一时间步骤的观察和反事实内部状态的组合来确定所述机器学习系统在下一时间步骤处的内部状态; 并且对于其他时间步骤,

通过使用所述内部状态更新子系统处理所述当前时间步骤的解码器输出、所述下一时间步骤的观察和所述当前时间步骤处的内部状态的组合来确定所述机器学习系统在所述下一时间步骤处的内部状态。

5. 根据权利要求1所述的方法, 还包括:

获得训练数据的数据集, 所述训练数据包括定义针对一系列时间步骤中的每个时间步骤的观察和动作的数据, 针对所述一系列时间步骤中的每个时间步骤, 所述训练数据已经通过如下来获得:

获得表征当前时间步骤的环境状态的观察; 以及

使用所述机器学习系统的所述动作选择子系统来处理所述机器学习系统在所述当前时间步骤处的所述内部状态, 以生成所述动作选择策略输出; 以及

针对一系列时间步骤中的每个时间步骤, 通过以下方式处理所述训练数据的数据集:

使用所述解码器神经网络处理所述机器学习系统在所述当前时间步骤处的内部状态, 以生成所述当前时间步骤的解码器输出;

并且其中处理所述训练数据的数据集包括:

通过使用内部状态更新子系统处理当前时间步骤的解码器输出、下一时间步骤的观察和反事实内部状态的组合来确定机器学习系统在下一时间步骤处的内部状态; 并且还包括:

使用目标函数训练所述机器学习系统, 所述目标函数取决于以下任一项:

i) 在a) 通过处理反事实内部状态确定的下一时间步骤处的代理控制系统的内部状态与b) 通过使用内部状态更新子系统处理当前时间步骤的解码器输出、下一时间步骤的观察和当前时间步骤处的代理控制系统的内部状态的组合确定的下一时间步骤处的代理控制系统的内部状态之间的差异; 或

ii) 在a) 根据代理控制系统在当前时间步骤处的内部状态确定的动作选择子系统的动作选择策略与b) 根据反事实内部状态确定的动作选择子系统的动作选择策略之间的差异。

6. 根据权利要求5所述的方法, 其中, 所述目标函数取决于所述代理控制系统的内部状态之间的差异, 并且其中, 通过使用所述内部状态更新子系统处理所述当前时间步骤的解码器输出、所述下一时间步骤的观察和所述反事实内部状态的组合来确定所述代理控制系统在所述下一时间步骤处的内部状态定义了根据干预时间表执行的干预, 并且其中, 所述干预时间表包括在每个时间步骤处执行所述干预。

7. 根据权利要求5所述的方法, 其中, 所述目标函数取决于所述动作选择策略之间的差异, 其中, 通过使用所述内部状态更新子系统处理所述当前时间步骤的解码器输出、所述下一时间步骤的观察和所述反事实内部状态的组合来确定所述代理控制系统在所述下一时间步骤处的内部状态定义了根据干预时间表执行的干预, 并且其中通过将时间步骤划分为可变持续时间的时间步骤块序列并且在每个时间步骤块处执行所述干预来定义所述干预

时间表。

8. 根据权利要求5或7所述的方法, 其中, 所述动作选择策略之间的差异包括由所述动作选择策略输出定义的分布差异的度量。

9. 根据权利要求5、7或8所述的方法, 其中, 所述动作选择策略之间的差异包括在多个时间步骤上的所述动作选择策略之间的差异之和。

10. 根据权利要求1-9中任一项所述的方法, 其中, 所述反事实内部状态包括所述机器学习系统在比所述当前时间步骤更早的时间步骤处的内部状态。

11. 根据权利要求10所述的方法, 其中, 所述反事实内部状态包括所述机器学习系统在第一时间步骤处的内部状态。

12. 根据权利要求1-11中任一项所述的方法, 其中, 所述当前时间步骤的解码器输出包括具有元素的向量, 每个元素表示所述机器学习系统的多个离散内部状态中的相应一个。

13. 根据权利要求1至12中任一项所述的方法, 其中, 所述当前时间步骤的解码器输出包括所述机器学习系统的内部状态的自然语言描述。

14. 根据权利要求1-13中任一项所述的方法, 还包括:

修改所述当前时间步骤的所述解码器输出, 以及

通过使用所述内部状态更新子系统处理所述当前时间步骤的所修改的解码器输出、所述下一时间步骤的所述观察和所述反事实内部状态的组合来确定所述机器学习系统在下一时间步骤处的内部状态。

15. 根据权利要求14所述的方法, 还包括通过以下方式控制所述代理在所述环境中的行为:

经由用户接口输出从当前时间步骤的解码器输出导出的信号;

从所述用户接口接收指令, 其中, 所述指令与所述任务在所述环境中的执行有关; 以及
从所述指令生成所修改的解码器输出。

16. 根据任一前述权利要求所述的方法, 其中, 所述动作选择子系统包括所述机器学习系统的动作选择策略神经网络。

17. 根据任一前述权利要求所述的方法, 其中, 所述内部状态更新子系统包括递归神经网络。

18. 根据任一前述权利要求所述的方法, 其中, 所述内部状态更新子系统包括变换器神经网络。

19. 根据任一前述权利要求所述的方法, 还包括通过反向传播解码器训练目标函数的梯度来训练所述解码器神经网络, 所述解码器训练目标函数的梯度取决于来自所述解码器输出的所述机器学习系统的内部状态的描述与所述环境的真实描述之间的差异。

20. 一种训练代理控制系统的计算机实现的方法, 所述代理控制系统用于监视在由所述代理控制系统控制的环境中操作以执行任务的代理的性能, 所述方法包括:

获得训练数据的数据集, 所述训练数据包括定义针对一系列时间步骤中的每个时间步骤的观察和动作的数据, 针对所述一系列时间步骤中的每个时间步骤, 所述训练数据已经通过如下来获得:

获得表征当前时间步骤的环境状态的观察; 以及

使用所述代理控制系统的动作选择子系统来处理所述代理控制系统在所述当前时间

步骤处的内部状态,以生成动作选择策略输出,其中,所述代理控制系统在所述当前时间步骤处的所述内部状态取决于所述当前时间步骤的所述观察;

所述方法还包括针对一系列时间步骤中的每个时间步骤,通过以下方式处理所述训练数据的数据集:

使用解码器神经网络处理所述代理控制系统在所述当前时间步骤处的所述内部状态,以生成所述当前时间步骤的解码器输出,其中,所述当前时间步骤的所述解码器输出描述所述代理控制系统在所述当前时间步骤处的所述内部状态;

通过使用内部状态更新子系统处理当前时间步骤的解码器输出、下一时间步骤的观察和反事实内部状态的组合来确定代理控制系统在下一时间步骤处的内部状态,其中反事实内部状态不同于代理控制系统在当前时间步骤处的内部状态;

使用目标函数训练所述解码器神经网络,所述目标函数取决于:

i) 在通过处理反事实内部状态确定的在下一时间步骤处的代理控制系统的内部状态与通过使用内部状态更新子系统处理当前时间步骤的解码器输出、下一时间步骤的观察和当前时间步骤处的代理控制系统的内部状态的组合确定的下一时间步骤处的代理控制系统的内部状态之间的差异;或

ii) 在根据代理控制系统在当前时间步骤处的内部状态确定的动作选择子系统的动作选择策略与根据反事实内部状态确定的动作选择子系统的动作选择策略之间的差异。

21. 根据权利要求20所述的方法,其中,所述动作选择策略之间的差异包括由所述动作选择策略输出定义的分布差异的度量。

22. 根据权利要求20或21所述的方法,其中,所述动作选择策略之间的差异包括在多个时间步骤上的所述动作选择策略之间的差异之和。

23. 一种机器学习系统,被配置为监视环境中执行任务的代理的控制,所述机器学习系统包括:

解码器神经网络,被配置为在时间步骤处处理所述机器学习系统的内部状态以生成所述时间步骤的解码器输出,其中所述时间步骤的解码器输出描述所述机器学习系统在所述时间步骤处的内部状态;

动作选择子系统,被配置为在时间步骤处处理所述机器学习系统的内部状态,以生成用于选择要由所述代理在所述时间步骤处执行的动作用的动作选择策略输出;以及

内部状态更新子系统,其被配置为处理当前时间步骤的解码器输出、下一时间步骤的观察、以及时间步骤的当前时间步骤处的机器学习系统的内部状态的组合,以生成更新的内部状态;以及

输出,用于提供从时间步骤的解码器输出导出的信号以用于监视环境中的代理的控制。

24. 根据权利要求23所述的系统,还包括用于控制所述代理在所述环境中的行为的用户接口;其中,所述系统被配置为:

经由所述用户接口输出从所述解码器输出导出的所述信号;

从所述用户接口接收指令,其中,所述指令与所述任务在所述环境中的执行相关;

从所述指令生成修改的解码器输出;以及

处理当前时间步骤的修改的解码器输出、下一时间步骤的观察以及时间步骤的当前时

间步骤处的机器学习系统的内部状态的组合,以生成更新的内部状态。

25.根据权利要求1至22中任一项所述的方法或根据权利要求23或24所述的系统,其中,所述代理是机械代理,所述环境是现实世界环境,所述观察来自感测所述现实世界环境的一个或多个传感器,并且所述动作控制所述机械代理在所述现实世界环境中动作以执行所述任务。

26.一种系统,包括一个或多个计算机和存储指令的一个或多个存储设备,所述指令在由所述一个或多个计算机执行时可操作以使所述一个或多个计算机执行根据权利要求1-22中任一项所述的相应方法的操作。

27.一种编码有指令的计算机存储介质,所述指令在由一个或多个计算机执行时使所述一个或多个计算机执行根据权利要求1-22中任一项所述的相应方法的操作。

具有反事实干预的机器学习系统

技术领域

[0001] 本说明书总体上涉及使用神经网络控制代理。

背景技术

[0002] 神经网络是采用一层或多层非线性单元来预测接收的输入的输出的机器学习模型。一些神经网络除了输出层之外还包括一个或多个隐藏层。每个隐藏层的输出被用作网络中的下一层(即,下一隐藏层或输出层)的输入。网络的每个层根据相应参数集的当前值从接收的输入生成输出。

发明内容

[0003] 本说明书涉及在一个或多个位置中的一个或多个计算机上实现为计算机程序的系统和方法,用于训练和使用机器学习系统来控制代理执行任务。机器学习代理控制系统使用反事实内部状态来训练,使得它可以提供输出,该输出以因果的方式解释系统的行为,例如,就其环境的方面而言,该环境的方面使得系统为代理选择特定动作。

[0004] 在一个方面,描述了一种训练机器学习代理控制系统(例如,强化学习系统)的计算机实现的方法,该机器学习代理控制系统被配置为监视环境中执行任务的代理的控制。

[0005] 机器学习系统被配置为,对于一系列时间步骤中的每个时间步骤,获得表征当前时间步骤的环境状态的观察。该系统还被配置为使用动作选择子系统来处理机器学习系统在当前时间步骤的内部状态,以生成动作选择策略输出。机器学习系统在当前时间步骤的内部状态取决于当前时间步骤的观察。该系统被配置为使用解码器神经网络处理机器学习系统在当前时间步骤的内部状态以生成解码器输出。解码器输出描述了机器学习系统在当前时间步骤的内部状态。系统响应于观察使用动作选择策略输出来选择要由代理在当前时间步骤执行的动作,并且使代理执行所选择的动作。系统还提供从当前时间步骤的解码器输出导出的信号作为用于监视环境中的代理的控制的监视信号。

[0006] 该方法包括通过使用内部状态更新子系统处理当前时间步骤的解码器输出、下一时间步骤的观察和反事实内部状态的组合来确定机器学习系统在下一时间步骤的内部状态,从而训练机器学习系统。反事实内部状态不同于机器学习系统在当前时间步骤处的内部状态。

[0007] 在一些实施方式中,反事实内部状态包括机器学习系统在比当前时间步骤更早的时间步骤处的内部状态,例如,机器学习系统在与环境交互的代理的情节的第一时间步骤处的内部状态,在该第一时间步骤期间代理尝试执行任务。

[0008] 在一些实施方式中,机器学习系统已经被训练;例如,上述过程可以用于微调系统。在一些实施方式中,使用先前存储的数据离线执行该过程,而不需要代理与环境进一步交互。

[0009] 因此,在另一方面,一种过程获得训练数据的数据集,该数据集包括定义针对一系列时间步骤中的每个时间步骤的观察和动作的数据,并且使用该数据集并基于一个或多个

反事实内部状态来训练机器学习系统。

[0010] 在机器学习(代理控制)系统已经被训练之后,监视信号(例如解码器输出)可以用于监视系统对环境中的代理的控制,以执行任务。例如,监视信号可以用于获得根据系统对环境的信念来解释系统/代理的行为的信息。可选地,可以修改解码器输出,以修改系统关于环境的信念,从而干预代理的控制。

[0011] 本说明书中描述的主题可以在特定实施例实现,以便实现以下优点中的一个或多个。

[0012] 所描述的技术可以提供具有输出的机器学习代理控制系统,该输出提供系统行为的因果解释。也即,输出中的信息是系统用于控制代理行为的信息,而不是仅与代理行为相关的信息。在实施方式中,这通过将输出反馈到机器学习系统中并且在训练期间使用反事实内部状态来确保输出被依赖来实现。

[0013] 在实施方式中,输出中的信息在语义上以环境的特征为基础,即,输出可以包括关于环境的特征的描述。输出的形式是灵活的,并且它可以产生各种不同形式的表示,例如类、图形和字符串,特别是自然语言描述。

[0014] 改变描述产生与改变的描述一致的行为,并且因此所描述的技术还使得能够进行干预以允许控制系统的行为,例如出于安全目的。改变的描述也可以是基于语义的方式,即人类可解释的,例如以自然语言的形式。以这种方式,可以引导代理,更具体地控制代理的机器学习系统,或者可以在安全关键应用中施加对代理行为的限制。例如,在机器人的情况下,可以指示系统关于要拾取或操纵的物体的位置,或者在速度限制为例如40mph的自动驾驶车辆的情况下,可以指示系统速度限制应该改为30mph。

[0015] 所描述的技术的一些实施方式使得能够离线分析机器学习代理控制系统的行为(例如,在仿真中)。然后,解释系统行为的输出可以用于可选地自动确定机器学习代理控制系统之后是否应该控制代理执行任务(例如,在现实世界中)。

[0016] 这些技术不限于特定类型的机器学习代理控制系统,并且扩展到具有许多参数的机器学习系统。例如,它们可以与使用模仿学习的机器学习代理控制系统一起使用,例如行为克隆、反向模仿学习或生成对抗模仿学习(GAIL,arXiv:1606.03476,Ho等人),或者与使用基于响应于所执行的动作而接收到的奖励的强化学习的机器学习代理控制系统一起使用。

[0017] 它们可以对机器学习系统的性能具有相对较小的影响,特别是如果离线应用,并且它们可以使得能够在没有许多小时的人工反向工程的情况下解释机器学习系统的行为。

[0018] 在附图和下面的描述中阐述了本说明书的主题的一个或多个实施例的细节。根据说明书、附图和权利要求,主题的其他特征、方面和优点将变得显而易见。

附图说明

[0019] 图1示出了机器学习代理控制系统的示例。

[0020] 图2A和图2B示出了机器学习系统中的处理的示意图。

[0021] 图3A和图3B示出了使用反事实内部状态的处理的示意图。

[0022] 图4是用于训练机器学习系统的第一示例过程的流程图。

[0023] 图5是用于训练机器学习系统的第二示例过程的流程图。

- [0024] 图6是用于监视机器学习系统的示例过程的流程图。
- [0025] 图7示出了经训练的机器学习系统的性能。
- [0026] 各个附图中相同的附图标记和名称指示相同的元件。

具体实施方式

[0027] 图1示出了用于代理控制的机器学习系统100的示例,该机器学习系统100可以被实现为一个或多个位置中的一个或多个计算机上的一个或多个计算机程序。机器学习系统100用于控制与环境106交互的代理104以执行任务,并且生成人类可解释的解码状态(z_t),其解释系统行为。在实施方式中,系统被训练为在控制代理102时依赖于这些状态,并且因此它们提供系统行为的因果解释(而不仅仅是与系统的行为相关联)。例如,如果这些状态中的一个例如由用户经由用户接口170改变,则系统的行为也被改变。

[0028] 机器学习系统100可以包括任何类型的代理控制系统,例如,使用诸如行为克隆和/或强化学习的模仿学习。尽管机器学习系统100可以被松散地称为强化学习系统,但是这应当被理解为包括例如模仿学习。此外,所描述的技术的实施方式允许代理在环境中起作用,但不需要代理在环境中起作用-例如,稍后描述的CST-MR和CST-PD变体可以完全在重放中执行,而不需要代理在现场环境中起作用(根据反事实策略)。

[0029] 机器学习系统100具有一个或多个输入,以从环境接收表征环境状态的数据,例如,来自环境的一个或多个传感器的数据。表征环境状态的数据在本文中被称为观察110。观察可以是多模态观察,例如静止或运动图像与自然语言的组合。机器学习系统100在一系列时间步骤中的每一个处获得环境104的观察110。

[0030] 在时间 t 的观察的表示稍后表示为 x_t ;它可以包括在一个或多个先前时间步骤处的观察历史的表示。例如,可以使用ResNet(残差神经网络)对图像观察进行编码;或者,在标记化和向量化之后,可以使用LSTM(长短期记忆)神经网络对例如自然语言的文本输入进行编码。多模态观察可以包括要执行的任务的自然语言描述,可选地在每次观察时重复。

[0031] 来自环境的数据还可以包括任务奖励。通常,任务奖励108由表征代理朝向任务目标的进展的标量数值(其可以是零)表示,并且可以基于环境中的任何事件或环境的方面。任务奖励可以随着任务进行或者仅在任务结束时被接收,例如以指示任务的成功完成。

[0032] 机器学习系统100通过在多个动作选择时间步骤中的每个步骤处处理观察110以选择要由代理104执行的动作102来控制代理。动作102可以是连续的或离散的;它可以包括要在时间步骤执行的多个单独或原始动作的集合,例如连续和离散动作的混合。在每个时间步骤处,该时间步骤处的环境状态取决于先前时间步骤处的环境状态和代理在先前时间步骤处执行的动作。由代理104执行所选择的动作通常使环境104转变到新状态。通过重复地使代理102在环境104中动作,系统100可以控制代理102完成指定的任务。

[0033] 机器学习系统100包括动作选择子系统130,其生成用于选择动作102的动作选择策略输出132。通常,任何类型的动作选择子系统可以用于动作选择子系统130。在一些实施方式中,但不一定,动作选择子系统包括动作选择策略神经网络。

[0034] 存在可以使用动作选择策略输出132来选择动作的许多方式。例如,动作选择策略输出132可以直接定义动作,例如,它可以包括用于定义动作的连续值的值,诸如扭矩或速度。作为另一示例,它可以定义(例如,参数化)连续或分类分布,从该分布可以选择定义动

作的值;或者它可以将分布定义为一组分数,一组可能动作中的每个动作一个分数。这样的分布可以用于选择动作,例如通过从分布中采样或通过选择具有最高概率的动作。

[0035] 机器学习系统100在时间 t 维持内部状态 m_t ,并且包括内部状态更新子系统120。内部状态更新子系统120被配置为在当前时间步骤 t 处理内部状态 m_t ,当前时间步骤的解码器输出 z_t (稍后描述),以及下一时间步骤的观察的表示 x_{t+1} ,以生成用于下一时间步骤的更新的内部状态 m_{t+1} 。当前时间步骤的内部状态 m_t 由动作选择子系统130处理,以生成用于在当前时间步骤选择动作102的动作选择策略输出132。

[0036] 本文描述的技术不依赖于内部状态更新子系统120的任何特定架构。作为一个示例,在一些实施方式中,内部状态更新子系统120包括递归神经网络,诸如LSTM神经网络。然后内部状态 m_t 可以包括递归神经网络的递归状态。作为另一示例,在一些实施方式中,内部状态更新子系统120包括变换器神经网络。通常,变换器网络可以被表征为神经网络,特别是自回归神经网络,其具有一系列自注意力神经网络层。自注意力神经网络层具有针对输入的每个元素的注意力层输入,并且被配置为在注意力层输入上应用注意力机制以生成针对输入的每个元素的注意力层输出;存在许多可能的注意力机制。当内部状态更新子系统120包括变换器神经网络时,内部状态 m_t 例如可以包括先前观察的嵌入的可变长度阵列 $x_{0:t}$ 。在这些实施方式中的一些实施方式中,但不一定,内部状态更新子系统120包括存储器。

[0037] 机器学习系统100还包括解码器神经网络140,其被配置为根据解码器神经网络参数处理在当前时间步骤的内部状态 m_t ,以生成当前时间步骤的解码器输出 z_t 。通常,解码器神经网络140可以具有任何合适的架构,并且可以包括例如一个或多个前馈神经网络层、一个或多个卷积神经网络层、一个或多个注意力神经网络层或一个或多个归一化层。在一个示例实施方式中,解码器神经网络(例如多层感知器(MLP))参数化分布 $q(z_t)$,其中采样 z_t 的值, $z_t \sim q(z_t)$ 。在另一示例中, z_t 由LSTM提供。

[0038] 训练机器学习系统100,使得解码器输出 z_t 描述在 m_t 的信息,这实际上用于确定动作选择策略输出132,而不是仅仅反映在 m_t 的信息内容。

[0039] 所述解码器输出 z_t 用于提供监视信号132,用于监视机器学习系统100对环境中的代理的控制,例如,作为有线或无线网络上的信号或作为来自本地或远程用户接口170的输出。监视信号132可以是解码器输出 z_t ,或者它可以从处理解码器输出 z_t 导出。在一些实施方式中,解码器输出 z_t 提供机器学习系统100的行为的自然语言解释。

[0040] 在实施方式中,解码器输出 z_t 以及监视信号132,每个都包括系统关于环境104的“信念”的描述,即系统控制代理所依赖的环境的假设状态的描述。在一些实施方式中,解码器输出 z_t 以语义上有意义的术语表示,如稍后进一步描述的。所述解码器输出 z_t 以及监视信号132可以被视为对系统的“内部语音”的窃听。在实施方式中,对于当前时间步骤的内部状态 m_t 和所述解码器输出 z_t ,由内部状态更新子系统120组合(例如级联)并联合处理。在必要时,解码器输出 z_t 可以将其转换成适合于与 m_t 级联的形式。

[0041] 作为一个说明,在一些实施方式中,用于当前时间步骤的解码器输出 z_t 包括向量(例如1-hot向量),所述向量具有各自表示机器学习系统100的多个离散内部状态(即,信念状态)中的相应一个的元素。例如,环境可以包括多个对象(τ 个),每个对象具有表示与对象相关的信念的相关联的1-hot向量,例如,识别对象的状态、配置、描述或位置的向量的每个

元素。分布 $q(z_t)$ 可以定义为独立分类分布的乘积 $q(z_t) = \prod_{i=1}^T q(z_t^i)$ 以及 z_t 可以通过从该分布采样来确定。

[0042] 作为另一个说明,在一些实施方式中,用于当前时间步骤的解码器输出包括机器学习系统的内部状态的自然语言描述,其表示机器学习系统关于环境的信念。解码器神经网络可以定义表示标记(诸如词块、单词、句子的部分或句子)的概率的一组或多组分数,其可以用于例如通过采样或通过选择最可能的输出来确定自然语言输出。例如,用于当前时间步骤的解码器输出, z_t 可以从由LSTM的输出定义的分布中采样。值 z_t 可以对应于自然语言中的句子,其可以根据LSTM输出从连续样本构建,其中每个连续样本表示来自预定标记词汇表的标记。

[0043] 机器学习代理控制系统100可以是维持内部状态的任何类型的机器学习系统,例如强化学习系统。例如,它可以是基于模型或无模型的强化学习系统。它可以基于时差学习(例如Q学习),或者基于直接学习策略(例如经由策略梯度),或者它可以使用另一种策略优化技术,诸如MP0(arXiv:1806.06920)或PP0(arXiv:1707.06347)。机器学习系统(例如,强化学习系统)可以实现参与者-评论者技术。机器学习系统可以是分布式系统。

[0044] 可以通过使用任何形式的强化学习目标函数并使用任何适当的梯度下降优化算法(例如,Adam)迭代地更新系统的可学习参数(例如,神经网络权重)来训练机器学习系统100。仅作为一个示例,强化学习目标函数可以包括V-轨迹更新分量(arXiv:1802.01561)。通常,机器学习系统100的训练可以由训练引擎150控制;在训练之后,这是不需要的。

[0045] 机器学习代理控制系统100(例如强化学习系统)可以在线训练,即通过环境106采取动作并基于接收到的奖励更新可学习参数,例如以鼓励接收到的奖励的累积度量(例如奖励或“回报”的时间折扣总和)的增加。此外或替代地,机器学习系统100可以离线训练,即使用先前收集的数据而不与环境进一步交互。离线训练数据可以存储在离线训练数据存储160中;作为示例,对于一系列时间步骤,它可以包括对环境106的观察、指示所执行的动作的数据、在下一时间步骤对环境的后续状态的观察以及可选地指示响应于动作而接收到的奖励的数据。离线训练数据可以通过与环境106的人和/或机器交互来生成。

[0046] 在所描述的技术的实施方式中,无论机器学习系统100是在线学习、离线学习还是在线学习和离线学习,机器学习系统100都被训练为当选择动作时依赖于解码器输出, z_t 。这是通过在训练期间的一个或多个或每个时间步骤处,通过使用内部状态更新子系统120处理当前时间步骤的解码器输出 (z_t), 下一个时间步骤的观察, x_{t+1} 以及反事实的内部状态, m_t 的组合来确定机器学习系统在下一个时间步骤处的内部状态而完成的。反事实的内部状态, m_t 是与所述机器学习系统在所述当前时间步骤的内部状态 (m_t) 不同的状态 (例如,它具有与当前内部状态相同的格式,但具有与当前内部状态不同的值)。

[0047] 例如,考虑状态更新函数 $m_{t+1}=f_{\theta}([x_{t+1}, z_t], m_t)$ 和动作选择策略 $\pi_t=h_{\theta}(m_t)$ 其中 θ 表示学习的参数。因为解码器输出 z_t 通过处理 m_t 获得,当执行更新时状态更新函数不需要依赖于的值 z_t ,因此不能保证 z_t 反映系统的因果信念状态。通过使用反事实内部状态训练系统至少一些时间步骤,即,使得 $m_{t+1}=f_{\theta}([x_{t+1}, z_t], m_t)$,系统必须依赖于值 z_t 达至少一些时间步骤以获得关于当前状态的信息。

[0048] 可以获得反事实内部状态 (m_t) 的一种方式是使用来自不同时间步骤 t' 的内部状态作为反事实的内部状态。然后可以将更新的内部状态表示为:

$m_{t+1}^{(t \leftarrow t')}$ = $f_{\theta}([x_{t+1}, z_t], m_{t'})$ 以及对应的策略 $\pi_t^{(t \leftarrow t')} = h_{\theta}(m_{t+1}^{(t \leftarrow t')})$ 。不同的时间步骤可以是任何其他时间步骤,例如较早的时间步骤。例如,在一些实施方式中,不同的时间步骤可以是第一时间步骤,即 $t' = 0$ (这相当于内部状态上的丢失,其然后可以是归零或随机状态)。

[0049] 在训练期间,可以执行干预以用反事实内部状态替换当前内部状态。这可以根据干预时间表来完成;例如随机地,或在每个时间步骤处,或针对可变持续时间的时间步骤块序列中的每一个。

[0050] 图2A是在处理观察 (x_t, x_{t+1}) 时执行的处理步骤的示意图,以生成动作选择策略输出 π_t, π_{t+1} 和解码器输出 z_t , 更新当前内部状态从 m_t 至 m_{t+1} 。图2B示出了这样的处理步骤的序列。

[0051] 图3A示出了在更新内部状态时,使用反事实内部状态 $m_{t'}$, 而不是 m_t 进行干预的处理步骤。图3B示出了这样的干预的一些示例,其中,从上到下, $(t=1 \leftarrow t=0)$, $(t=3 \leftarrow t=0)$ 以及 $(t=3 \leftarrow t=1)$ 例如,当 $(t=3 \leftarrow t=1)$ 内部状态 m_4 还原为 m_1 以及 z_3 而不是 m_3 必须通知系统当前状态。

[0052] 在一些实施方式中,机器学习系统100包括强化学习系统,并且在没有修改的情况下使用强化学习目标函数(例如,如上所述)来训练。在这种情况下,可以训练系统以在执行干预时如前所述最大化回报。

[0053] 此外或替代地,目标函数可以包括一个或多个其他部件或由一个或多个其他部件组成。例如,目标函数可以包括鼓励在下一时间步骤准确重建内部状态的项,例如,项

$$\mathcal{L}_M = \left\| m_{t+1}^{(t \leftarrow t')} - m_{t+1} \right\|^2。$$

[0054] 作为另一示例,目标函数可以包括鼓励在干预下的动作选择策略匹配没有干预的策略的项,例如,项 $\mathcal{L}_P = \sum_{\Delta t > 0} \gamma(\Delta t) \cdot D_{KL}(\pi_{t+\Delta t} || \pi_{t+\Delta t}^{(t \leftarrow t')})$, 其中 Δt 定义一个时间步骤块, $\gamma(\cdot)$ 定义折扣因子(其可以是单位), $\pi_{t+\Delta t}^{(t \leftarrow t')}$ 以及 $\pi_{t+\Delta t}$ 在有干预和没有干预的情况下的相应策略,以及 $D_{KL}(\cdot)$ 是KL-散度。

[0055] 在一些实施方式中,当使用反事实内部状态参数进行训练时,可以限制更新,例如,使得动作选择子系统130的可学习参数不被更新,以限制对动作选择策略的中断。在目标函数包括上述一个或多个项的情况下,训练可以完全离线执行,并且目标函数不需要包括强化学习组件。也即,机器学习系统100可以使用仅包括 $\mathcal{L}_M$ 或 $\mathcal{L}_P$ 项的总损失来离线训练。这可以将策略学习与训练系统的用于提供系统行为的因果解释的部分分开,从而限制这对动作选择策略的任何潜在影响。

[0056] 在一些实施方式中,用户接口170可以用于修改解码器输出 z_t , 例如,上述1-hot向量或自然语言表示,在其用于控制代理之前,即在由内部状态更新子系统120处理之前。

[0057] 解码器输出描述了环境106(它反映了系统对环境的信念),因此可以向系统给予关于环境的(反事实的)信息,例如,以施加与安全相关的假设,诸如速度限制为30mph,而实际上速度限制为40mph。因为机器学习系统100依赖于解码器输出 z_t 以选择动作,来自用户接口的(反事实)信息可以改变所选择的动作和代理的行为。

[0058] 图4是用于训练机器学习系统的第一示例过程的流程图。图4的过程可以由一个或多个位置中的一个或多个计算机实现,例如,在强化学习神经网络系统100上。可以针对一系列时间步骤中的每一个执行该过程。图4的步骤不需要以所示的顺序执行。

[0059] 在步骤400,该过程获得表征当前时间步骤 t 的环境状态的观察。此外,在当前时间步骤,机器学习系统的内部状态, m_t ,使用动作选择子系统130处理以生成动作选择策略输出132(步骤402)。所述机器学习系统的内部状态, m_t ,还使用解码器神经网络140进行处理,以生成当前时间步骤的解码器输出, z_t (步骤404),其可以被输出。在实施方式中,该方法提供从当前时间步骤的解码器输出导出的信号作为用于监视环境中机器学习系统对代理的控制的监视信号(步骤406)。使用动作选择策略输出132选择动作并由代理执行动作(步骤408),并且接收下一观察和奖励(步骤410)。

[0060] 该过程通过使用内部状态更新子系统120处理当前时间步骤的解码器输出、下一时间步骤的观察和内部状态的组合来确定机器学习系统在下一时间步骤的内部状态。在一些时间步骤处,处理当前时间步骤处的解码器输出、下一时间步骤处的观察和当前时间步骤处的内部状态的组合(步骤412)。在一些时间步骤处,例如在每个干预时间步骤处,处理当前时间步骤处的解码器输出、下一时间步骤处的观察和反事实内部状态的组合(步骤412)。

[0061] 干预时间步骤是当根据干预时间表执行干预时的时间步骤。干预时间步骤可以是任何时间步骤或每个时间步骤(在实施方式中,在第一时间步骤之后)。在一些实施方式中,干预时间表包括随机选择干预的时间步骤。

[0062] 在实施方式中,机器学习系统100包括强化学习系统100,并且基于观察、所选择的动作以及在实施方式中接收的奖励使用强化学习目标函数进行训练;执行动作和接收奖励时在线或者离线(步骤414)。

[0063] 图5是用于训练机器学习系统(特别是代理控制系统,例如强化学习系统)的第二示例过程的流程图,特别是用于训练如上所述的解码器神经网络(和内部状态更新系统)。图5的过程可以由一个或多个位置中的一个或多个计算机实现,并且可以离线执行,即使用先前存储的数据而不需要代理进一步与环境交互。例如,图5的过程可以用于训练机器学习系统100,可选地在在线训练系统之后,或者它可以用于训练另一代理控制系统。

[0064] 在步骤500处,该过程获得训练数据的数据集,该训练数据包括定义针对一系列时间步骤中的每个时间步骤的观察和动作的数据。训练数据可以包括奖励数据,但不需要(例如,如果使用行为克隆)。在一些实施方式中,已经使用机器学习系统100或另一代理控制系统获得了训练数据;在一些实施方式中,它可能已经从人的动作获得。

[0065] 作为示例,可以在训练机器学习系统100的同时获得训练数据的数据集,而不使用任何反事实内部状态。更详细地,在一些实施方式中,已经通过以下方式获得训练数据:对于一系列时间步骤中的每一个,获得表征当前时间步骤的环境状态的观察,并且使用动作选择子系统(例如动作选择子系统130)处理当前时间步骤处的系统的内部状态,以生成用于控制代理的动作选择策略输出。

[0066] 针对一系列时间步骤中的每一个(例如,对应于训练数据的数据集中的时间步骤),使用强化学习或其他代理控制系统来处理训练数据的数据集。这涉及使用解码器神经网络处理系统在当前时间步骤的内部状态,以生成当前时间步骤的解码器输出(步骤502)。

[0067] 通过使用内部状态更新子系统处理当前时间步骤的解码器输出、下一时间步骤的观察和反事实内部状态的组合来确定系统在下一时间步骤的内部状态(步骤504)。图5的过程的一些实现在每个时间步骤处处理反事实内部状态。在图5的过程的一些实施方式中,执行该过程的一系列时间步骤包括时间步骤块;连续的时间步骤块可以具有不同的持续时间。

[0068] 使用目标函数来训练强化学习神经网络系统100或其他代理控制系统,该目标函数鼓励系统使用当前时间步骤的解码器输出来重建下一时间步骤的内部状态,或者鼓励系统恢复真实的当前动作选择策略,或者两者(步骤506)。

[0069] 鼓励系统使用当前时间步骤的解码器输出来重建下一时间步骤的内部状态的目标函数可以是取决于如下的a)与b)之间的差异的目标函数,其中,a)通过处理反事实内部状态确定的下一时间步骤的代理控制系统的内部状态,以及b)通过使用内部状态更新子系统处理当前时间步骤的解码器输出、下一时间步骤的观察,以及代理控制系统在当前时间步骤的内部状态的组合而确定的下一时间步骤的代理控制系统的内部状态。可以针对每个时间步骤评估这样的目标函数。

[0070] 鼓励系统恢复真实的当前动作选择策略的目标函数可以是取决于如下的a)与b)之间的差异的目标函数,其中,a)根据代理控制系统在当前时间步骤的内部状态确定的动作选择子系统的动作选择策略,以及b)根据反事实内部状态确定的动作选择子系统的动作选择策略。可以针对时间步骤块来评估这样的目标函数。动作选择策略之间的差异可以通过由动作选择策略输出定义的分布差异的度量来确定,例如,确定为在多个时间步骤的轨迹上的动作选择策略之间的差异之和。

[0071] 如前所述,在上述系统的一些实施方式中,解码器输出 z_t 用语义上有意义的术语表示。虽然不是必需的,但是这可以通过基于对环境的真实知识训练解码器输出 z_t 来促进。例如,在代理的任务是移动到指定对象的位置的情况下,真实知识可以是指定对象的位置的识别,例如,作为1-hot向量或以自然语言,诸如“<指定对象>在<对象位置>”。该信息与任务相关(在该示例中,是任务所需的),因此与系统执行任务需要使用的信息兼容。解码器神经网络140可以基于这样的真实知识例如使用监督学习来训练。这样的真实知识可能仅对于几个训练示例是需要的。

[0072] 因此,图4和图5的过程的一些实施方式可以包括使用基于解码器训练目标函数的监督训练来训练解码器神经网络,即更新解码器神经网络的可学习参数,诸如权重。解码器训练目标函数可以是任何合适的损失函数,其取决于来自解码器输出(针对当前时间步骤)的机器学习系统的内部状态(在当前时间步骤)的描述与环境的真实描述(针对当前时间步骤)之间的差异。

[0073] 图6是用于监视并且可选地干预对执行任务的环境中的代理的控制的示例过程的流程图。图6的过程可以由一个或多个位置中的一个或多个计算机实现,例如,在系统已经被训练之后通过如上所述的强化学习或其他代理控制系统实现。图6的步骤可以以与所示顺序不同的顺序执行。

[0074] 在每个时间步骤,该过程获得表征环境状态的观察(步骤600)。一个或多个观察可以包括对要执行的任务的描述。内部状态更新子系统120结合当前解码器输出 z_t 以及用于下一时间步骤的观察的表示, x_{t+1} ,来处理当前时间步骤 t 的内部状态, m_t 以生成更新的内部

状态, m_{t+1} (步骤602)。动作选择子系统120处理更新的内部状态, m_{t+1} , 以生成用于选择要由代理104执行的动作的下一时间步骤的动作选择策略输出132 (步骤604)。解码器神经网络140处理更新的内部状态, m_{t+1} , 以生成用于下一时间步骤的所述解码器输出, z_{t+1} , 以例如经由用户接口170提供监视环境中的代理的控制的监视信号142 (步骤606)。

[0075] 在一些实施方式中, 用户接口170或另一用户接口用于控制代理在环境中的行为。从解码器输出导出的信号 (例如, 用于当前时间步骤) 可以经由如前所述的用户接口提供给用户。该过程还可以包括从用户接口接收与环境中的任务的执行相关的指令, 例如, 以通过指定环境的一个或多个方面或特性来重新定义机器学习系统关于环境的信念 (步骤608)。然后, 从指令生成修改的解码器输出, 例如, 通过相应地修改自然语言或1-hot表示 (步骤610)。然后在当前时间步骤中使用修改后的解码器输出, 而不是当前时间步骤的真实解码器输出 (步骤602)。

[0076] 内部状态更新子系统120可以处理当前时间步骤的修改的解码器输出、下一时间步骤的观察和机器学习系统在当前时间步骤的内部状态的组合, 以生成更新的内部状态 (步骤612)。以这种方式, 修改机器学习系统关于环境的信念以引导代理的行为。与提供指令作为输入以控制机器学习系统的方法不同, 使用所描述的技术, 机器学习系统致力于跟随来自系统的输出, 即解码器输出 (z_t), 其被修改并反馈到系统中, 具体是反馈到内部状态更新单元120中。

[0077] 在图6的过程的变型中, 可以从先前在使用机器学习系统100选择代理的动作时收集的数据离线获得环境的观察, 并且该过程用于基于这些来分析和解释代理的行为。

[0078] 图7示出了上述技术的一些实施方式的性能。在图7中, CST-RL指的是图4的过程的实施方式, CST-MR指的是图5的过程的实施方式, 其中总损失包括 $\mathcal{L}_M$ 项, 而CST-PD是指图5的过程的实施方式, 其中总损失包括 $\mathcal{L}_P$ 项。Ord-Dec是指一种技术, 其中 z_t 不用于更新系统的内部状态, 而仅提供输出; 并且RR-Dec是指一种技术, 其中 z_t 用于更新系统的内部状态, 但是其中不使用反事实内部状态。当考虑相关忠实性时, 不同的技术表现得类似, 但是只有使用反事实内部状态的技术才表现出因果忠实性, 即解码器输出 (z_t) 中的信息以及监视信号142对由系统选择的动作具有因果影响。

[0079] 本文描述的技术可广泛应用, 并且不限于一个特定实施方式。然而, 为了说明的目的, 下面描述了少量的示例实施方式。

[0080] 在一些实施方式中, 代理是机械代理, 环境是现实世界环境, 观察来自感测现实世界环境的一个或多个传感器, 并且动作控制在现实世界环境中动作的机械代理以执行任务。

[0081] 更详细地, 在一些实施方式中, 环境是现实世界环境, 代理是与现实世界环境交互的机械代理, 例如, 在环境中操作或导航通过环境的机器人或自主或半自主陆地、空中或海上运载工具, 并且动作是由机械代理在现实世界环境中采取以执行任务的动作。例如, 代理可以是与环境交互以完成特定任务的机器人, 例如, 定位环境中的感兴趣对象或将感兴趣对象移动到环境中的指定位置或导航到环境中的指定目的地。

[0082] 在这些实施方式中, 观察可以包括例如以下中的一个或多个: 图像、对象位置数据和传感器数据, 以在代理与环境交互时捕获观察, 例如来自图像、距离或位置传感器或来自致动器的传感器数据。

[0083] 例如,在机器人的情况下,观察可以包括表征机器人的当前状态的数据,例如,以下中的一个或多个:关节位置、关节速度、关节力、扭矩或加速度,例如,重力补偿扭矩反馈,以及由机器人保持的物品的全局或相对姿势。在机器人或其他机械代理或运载工具的情况下,观察可以类似地包括代理的一个或多个部分的位置、线性或角速度、力、扭矩或加速度以及全局或相对姿势中的一个或多个。观察可以在1、2或3维中定义,并且可以是绝对和/或相对观察。观察还可以包括例如感测到的电子信号,诸如马达电流或温度信号;和/或例如来自相机或激光雷达传感器的图像或视频数据,例如,来自代理的传感器的数据或来自与环境中的代理分开定位的传感器的数据。

[0084] 在这些实施方式中,动作可以是控制信号,用于控制机器人或其他机械代理,例如,用于机器人的关节的扭矩或更高级别的控制命令,或者用于控制自主或半自主的陆地、空中、海上运载工具,例如,到控制表面或其他控制元件(例如,运载工具的转向控制元件)的扭矩,或更高级别的控制命令。控制信号可以包括例如机器人的一个或多个关节或另一机械代理的部件的位置、速度或力/扭矩/加速度数据。控制信号还可以或替代地包括电子控制数据,诸如马达控制数据,或者更一般地,用于控制环境内的一个或多个电子设备的数据,该一个或多个电子设备的控制对环境的观察状态有影响。例如,在自主或半自主陆地或空中或海上运载工具的情况下,控制信号可以定义用于控制导航(例如,转向)和移动(例如,运载工具的制动和/或加速)的动作。

[0085] 在一些实施方式中,环境是上述现实世界环境的仿真,并且代理被实现为与仿真环境交互的一个或多个计算机。例如,仿真环境可以是机器人或运载工具的仿真,并且机器学习系统可以在仿真上被训练,然后一旦被训练,就在现实世界中使用。

[0086] 在一些实施方式中,环境是用于制造产品(诸如化学、生物或机械产品或食物产品)的现实世界制造环境。如本文所用,“制造”产品还包括精制起始材料以产生产品,或处理起始材料,例如以除去污染物,以产生清洁或再循环的产品。制造工厂可以包括多个制造单元,例如用于化学或生物物质的容器,或用于处理固体或其他材料的机器,例如机器人。制造单元被配置成使得产品的中间版本或部件在产品的制造期间在制造单元之间可移动,例如经由管道或机械输送。如本文所使用的,产品的制造还包括通过厨房机器人制造食品。

[0087] 代理可以包括电子代理,该电子代理被配置为控制操作以制造产品的制造单元或诸如机器人的机器。也即,代理可以包括被配置为控制化学、生物或机械产品的制造的控制系统。例如,控制系统可以被配置为控制制造单元或机器中的一个或多个,或者控制产品的中间版本或部件在制造单元或机器之间的移动。

[0088] 作为一个示例,由代理执行的任务可以包括制造产品或其中间版本或组件的任务。作为另一示例,由代理执行的任务可以包括控制(例如,最小化)资源的使用的任务,诸如控制电力消耗或水消耗或制造过程中使用的任何材料或消耗品的消耗的任务。

[0089] 动作可以包括控制动作,以控制使用机器或制造单元来处理固体或液体材料以制造产品或其中间体或部件,或者控制产品的中间版本或部件在制造环境内(例如,在制造单元或机器之间)的移动。通常,动作可以是对观察的环境状态有影响的任何动作,例如,被配置为调整下面描述的任何感测参数的动作。这些可以包括调整制造单元的物理或化学条件的动作,或者控制机器的机械部件或机器人的关节的移動的动作。动作可以包括在制造单元或机器上施加操作条件的动作,或者导致改变设置以调整、控制或打开或关闭制造单元

或机器的操作的动作。

[0090] 奖励或回报可以涉及任务的执行的度量。例如,在要制造产品的任务的情况下,度量可以包括所制造的产品数量、产品的质量、产品的生产速度或执行制造任务的物理成本的度量,例如,用于执行任务的能量、材料或其他资源数量的度量。在要控制使用资源的任务的情况下,度量可以包括资源使用的任何度量。

[0091] 通常,环境状态的观察可以包括表示设备的电子和/或机械项目的功能的任何电子信号。例如,环境状态的表示可以从由感测制造环境的状态的传感器(例如,感测制造单元或机器的状态或配置的传感器,或者感测材料在制造单元或机器之间的移动的传感器)进行的观察中导出。作为一些示例,这样的传感器可以被配置为感测机械运动或力、压力、温度;电气条件,如电流、电压、频率、阻抗;一种或多种材料的数量、水平、流动/移动速率或流动/移动路径;物理或化学条件,例如物理状态、形状或构型或化学状态如pH;单元或机器的配置,诸如单元或机器的机械配置或阀门配置;图像或视频传感器,用于捕获制造单元或机器或运动的图像或视频观察;或任何其他适当类型的传感器。在诸如机器人的机器的情况下,来自传感器的观察可以包括对机器的一个或多个部件的位置、线性或角速度、力、扭矩或加速度或姿势的观察,例如,表征机器或机器人或由机器或机器人保持或处理的物品的当前状态的数据。观察还可以包括例如感测到的电子信号,诸如马达电流或温度信号,或者例如来自相机或激光雷达传感器的图像或视频数据。诸如这些的传感器可以是环境中的代理的一部分或与环境中的代理分开定位。

[0092] 在一些实施方式中,环境是包括多项电子设备的服务设施的现实世界环境,诸如服务器群或数据中心,例如电信数据中心,或用于存储或处理数据的计算机数据中心,或任何服务设施。服务设施还可以包括控制设备项目的操作环境的辅助控制设备,例如环境控制设备,诸如温度控制,例如冷却设备,或气流控制或空调设备。任务可以包括控制(例如,最小化)资源使用的任务,诸如控制电力消耗或水消耗的任务。代理可以包括电子代理,该电子代理被配置为控制设备项目的操作,或者控制辅助(例如,环境)控制设备的操作。

[0093] 通常,动作可以是对观察到的环境状态有影响的任何动作,例如,被配置为调整下面描述的任何感测参数的动作。这些可以包括控制设备项或辅助控制设备项或对其施加操作条件的动作,例如,导致改变设置以调节、控制或打开或关闭设备项或辅助控制设备项的操作的动作。

[0094] 通常,环境状态的观察可以包括表示设施或设施中的设备的功能的任何电子信号。例如,环境状态的表示可以从由感测设施的物理环境的状态的任何传感器进行的观察或由感测一个或多个设备项目或一个或多个辅助控制设备项目的状态的任何传感器进行的观察中导出。这些包括被配置为感测诸如电流、电压、功率或能量的电气条件的传感器;设施的温度;设施内或设施的冷却系统内的流体流量、温度或压力;或物理设施配置,诸如通风口是否打开。

[0095] 奖励或回报可以涉及任务的执行的度量。例如,在控制(例如,最小化)资源使用的任务(诸如控制电力或水的使用的任务)的情况下,度量可以包括资源使用的任何度量。

[0096] 在一些实施方式中,环境是发电设施(例如,诸如太阳能发电场或风电场的可再生发电设施)的现实世界环境。任务可以包括控制由设施生成的电力的控制任务,例如以控制向配电网的电力输送,例如以满足需求或降低电网的元件之间不匹配的风险,或最大化由

设施生成的电力。所述代理可以包括电子代理,所述电子代理被配置为控制所述设施的电力生成或将所生成的电力耦合到所述电网中。动作可以包括控制发电机的电气或机械配置的动作,诸如一个或多个可再生发电元件的电气或机械配置,例如控制风力涡轮机或一个或多个太阳能电池板或镜子的配置,或旋转发电机的电气或机械配置。机械控制动作可以例如包括控制能量输入到电能输出的转换的动作,例如能量输入到电能输出的转换效率或耦合程度。电控制动作可以例如包括控制所生成的电力的电压、电流、频率或相位中的一个或多个的动作。

[0097] 奖励或回报可以涉及任务的执行的度量。例如,在控制向配电网输送电力的任务的情况下,度量可以涉及所传递的电力的测量,或者涉及发电设施与电网之间的电失配的测量,诸如电压、电流、频率或相位失配,或者涉及发电设施中的电力或能量损耗的测量。在最大化到配电网的电力输送的任务的情况下,度量可以涉及传递到电网的电力或能量的测量,或者涉及发电设施中的电力或能量损耗的测量。

[0098] 通常,环境状态的观察可以包括表示发电设施中的发电装备的电气或机械功能的任何电子信号。例如,环境状态的表示可以从由感测发电设施中正在发电的设备的物理或电气状态、或这种设备的物理环境、或支持发电设备的辅助设备的状况的任何传感器进行的观察中导出。这样的传感器可以包括被配置为感测设备的电气状况(诸如电流、电压、功率或能量)的传感器;物理环境的温度或冷却;流体流动;或设备的物理配置;以及例如从本地或远程传感器观察电网的电气状况。环境状态的观察还可以包括关于发电设备的未来操作状况的一个或多个预测,诸如未来风水平或太阳辐照度的预测或电网的未来电气状况的预测。

[0099] 作为另一个实例,环境可以是化学合成或蛋白质折叠环境,使得每个状态是蛋白质链或一种或多种中间体或前体化学物质的相应状态,并且代理是用于确定如何折叠蛋白质链或合成化学物质的计算机系统。在该实例中,作用是用于折叠蛋白质链的可能的折叠作用或用于组装前体化学物质/中间体的作用,并且要实现的结果可以包括例如折叠蛋白质,使得蛋白质是稳定的并且使得它实现特定的生物学功能或为化学物质提供有效的合成途径。作为另一个实例,代理可以是机械代理,其执行或控制由系统自动选择的蛋白质折叠动作或化学合成步骤而无需人为交互。观察可以包括蛋白质或化学物质/中间体/前体的状态的直接或间接观察和/或可以源自仿真。

[0100] 以类似的方式,环境可以是药物设计环境,使得每个状态是潜在的药剂学药物的相应状态,并且代理是用于确定药剂学药物的元素和/或药剂学药物的合成途径的计算机系统。药物/合成可以基于源自药物靶标的奖励来设计,例如在仿真中。作为另一个实例,代理可以是执行或控制药物合成的机械代理。

[0101] 在一些进一步的应用中,环境是现实世界环境,并且代理管理跨计算资源(例如,在移动设备上和/或在数据中心中)的任务分配。在这些实施方式中,动作可以包括将任务分配给特定计算资源。

[0102] 因此,在一些应用中,代理是软件代理,并且环境是现实世界计算环境。在一个示例中,代理管理跨计算资源(例如,在移动设备上和/或在数据中心中)的任务分配。在这些应用中,观察可以包括对计算资源(诸如计算和/或存储器容量)或互联网可访问资源的观察;并且动作可以包括将任务分配给特定计算资源。奖励可以被配置为最大化或最小化以

下中的一个或多个:计算资源的利用率、电力、带宽和计算速度。

[0103] 在另一示例中,软件代理管理例如由一个或多个现实世界服务器对连续到达的作业的队列的处理。观察可以包括对连续作业的离开时间,或连续作业的离开之间的时间间隔,或服务器处理每个作业所花费的时间(例如,连续作业的时间范围的开始和结束,或到达时间,或到达之间的时间间隔,或表征作业类型的数据)的观察。动作可以包括将特定作业分配给特定计算资源的动作;奖励可以被配置为最小化总体排队或处理时间或一个或多个单独作业的排队或处理时间,或者通常基于观察来优化任何度量。

[0104] 作为另一示例,环境可以包括现实世界的计算机系统或网络,观察可以包括表征计算机系统或网络的操作的任何观察,由软件代理执行的动作可以包括控制操作的动作,例如以限制或校正例如由于存在病毒或其他安全漏洞而导致的异常或不期望的操作,并且奖励可以包括表征计算机系统或网络的期望操作的任何度量。

[0105] 在一些应用中,环境是现实世界计算环境,并且软件代理管理跨计算资源(例如,在移动设备上和/或在数据中心中)的任务/作业的分布。在这些实施方式中,观察可以包括与计算资源在处理任务/作业时的操作相关的观察,动作可以包括将任务/作业分配给特定计算资源,并且奖励可以涉及使用计算资源处理任务/作业的一个或多个度量,例如,计算资源、带宽或电力的使用的度量,或处理时间的度量,或数值精度,或与计算资源之间的期望负载均衡相关的一个或多个度量。

[0106] 在一些应用中,环境是数据分组通信网络环境,并且代理是用于通过通信网络路由数据分组的路由器的一部分。动作可以包括数据分组路由动作,并且观察可以包括例如路由表的观察,该路由表包括路由度量,诸如路由路径长度、带宽、负载、跳数、路径成本、延迟、最大传输单元(MTU)和可靠性的度量。任务相关奖励可以关于路由度量中的一个或多个来定义,即被配置为最大化路由度量中的一个或多个。

[0107] 在一些其他应用中,环境是互联网或移动通信环境,并且代理是管理用户的个性化推荐的软件代理。观察可以包括用户采取的先前动作,例如表征这些的特征;动作可以包括向用户推荐诸如内容项的各项的动作。任务相关奖励可以被配置为可选地在时间跨度内最大化以下中的一个或多个:用户将有利地响应被推荐(内容)项的估计可能性、一个或多个推荐项的适合性/不适合性、推荐各项的成本以及用户接收到的推荐的数量。

[0108] 作为另一示例,动作可以包括呈现广告,观察可以包括广告印象或点进计数或点进率,并且奖励可以表征一个或多个用户对项目或内容的先前选择。

[0109] 在一些情况下,观察可以包括由第三方(例如,代理的操作者)提供给代理的文本或口头指令。例如,代理可以是自主车辆,并且自主车辆的用户可以向代理提供文本或口头指令(例如,以导航到特定位置)。

[0110] 环境可以是仿真环境。通常,在仿真环境的情况下,观察可以包括先前描述的观察或观察类型中的一个或多个的仿真版本,并且动作可以包括先前描述的动作或动作类型中的一个或多个的仿真版本。例如,仿真环境可以是运动仿真环境,例如驾驶仿真或飞行仿真,并且代理可以通过运动仿真导航的仿真运载工具。在这些实施方式中,动作可以是用于控制仿真用户或仿真运载工具的控制输入。通常,代理可以被实现为与仿真环境交互的一个或多个计算机。

[0111] 仿真环境可以是特定现实世界环境和代理的仿真。例如,系统可以用于在系统的

训练或评估期间选择仿真环境中的动作,并且在训练或评估或两者完成之后,可以部署用于控制作为仿真对象的特定现实世界环境中的现实世界代理。这可以避免对现实世界环境或现实世界代理的不必要的磨损和损坏,并且可以允许在现实世界环境中很少发生或难以或不安全地重新创建的情况下训练或评估系统。在这种情况下,仿真环境的观察与现实世界环境相关,并且仿真环境中的所选动作与将由机械代理在现实世界环境中执行的动作相关。

[0112] 可选地,在任何上述实施方式中,在任何给定时间步骤的观察可以包括来自先前时间步骤的数据,其可以有益于表征环境,例如,在先前时间步骤执行的动作、在先前时间步骤接收的奖励或两者。

[0113] 本说明书结合系统和计算机程序组件使用术语“配置”。对于被配置为执行特定操作或动作的一个或多个计算机的系统,意味着系统已经在其上安装了软件、固件、硬件或它们的组合,这些软件、固件、硬件或它们的组合在操作中使得系统执行操作或动作。对于被配置为执行特定操作或动作的一个或多个计算机程序,意味着一个或多个程序包括当由数据处理装置执行时使装置执行操作或动作的指令。

[0114] 本说明书中描述的主题和功能操作的实施例可以在数字电子电路、有形体现的计算机软件或固件、计算机硬件(包括本说明书中公开的结构及其结构等同物)或它们中的一个或多个的组合中实现。本说明书中描述的主题的实施例可以实现为一个或多个计算机程序,即,在有形非暂时性存储介质上编码的计算机程序指令的一个或多个模块,用于由数据处理装置执行或控制数据处理装置的操作。计算机存储介质可以是机器可读存储设备、机器可读存储基板、随机或串行存取存储器设备、或它们中的一个或多个的组合。可替代地或另外地,程序指令可以被编码在人工生成的传播信号上,例如,机器生成的电、光或电磁信号,其被生成以对信息进行编码以便传输到合适的接收器装置以供数据处理装置执行。

[0115] 术语“数据处理装置”是指数据处理硬件,并且包含用于处理数据的所有种类的装置、设备和机器,例如包括可编程处理器、计算机或多个处理器或计算机。该装置还可以是或进一步包括专用逻辑电路,例如FPGA(现场可编程门阵列)或ASIC(专用集成电路)。除了硬件之外,该装置可以可选地包括为计算机程序创建执行环境的代码,例如,构成处理器固件、协议栈、数据库管理系统、操作系统或它们中的一个或多个的组的代码。

[0116] 计算机程序(其也可以被称为或描述为程序、软件、软件应用、app、模块、软件模块、脚本或代码)可以以任何形式的编程语言编写,包括编译或解释语言,或者声明或过程语言;并且它可以以任何形式部署,包括作为独立程序或作为模块、组件、子例程或适合在计算环境中使用的其他单元。程序可以但不必对应于文件系统中的文件。程序可以存储在保存其他程序或数据的文件的一部分中,例如,存储在标记语言文档中的一个或多个脚本,存储在专用于所讨论的程序的单个文件中,或者存储在多个协调文件中,例如,存储一个或多个模块、子程序或代码部分的文件。计算机程序可以被部署为在一个计算机上或在位于一个站点或分布在多个站点上并通过数据通信网络互连的多个计算机上执行。

[0117] 在本说明书中,术语“数据库”广泛地用于指代任何数据集合:数据不需要以任何特定方式结构化,或者根本不需要结构化,并且它可以存储在一个或多个位置中的存储设备上。因此,例如,索引数据库可以包括多个数据集合,每个数据集合可以被不同地组织和访问。

[0118] 类似地,在本说明书中,术语“引擎”广泛地用于指代被编程为执行一个或多个特定功能的基于软件的系统、子系统或过程。通常,引擎将被实现为安装在一个或多个位置中的一个或多个计算机上的一个或多个软件模块或组件。在一些情况下,一个或多个计算机将专用于特定引擎;在其他情况下,可以在同一台或多台计算机上安装和运行多个引擎。

[0119] 本说明书中描述的过程和逻辑流程可以由执行一个或多个计算机程序的一个或多个可编程计算机执行,以通过对输入数据进行操作并生成输出来执行功能。过程和逻辑流程也可以由专用逻辑电路(例如,FPGA或ASIC)执行,或者由专用逻辑电路和一个或多个编程计算机的组合执行。

[0120] 适合于执行计算机程序的计算机可以基于通用或专用微处理器或两者,或任何其他类型的中央处理单元。通常,中央处理单元将从只读存储器或随机存取存储器或两者接收指令和数据。计算机的元件是用于执行或运行指令的中央处理单元和用于存储指令和数据的一个或多个存储器设备。中央处理单元和存储器可以由专用逻辑电路补充或并入专用逻辑电路中。通常,计算机还将包括用于存储数据的一个或多个大容量存储设备(例如磁盘、磁光盘或光盘),或者可操作地耦合以从其接收数据或向其传输数据或两者。然而,计算机不需要具有这样的设备。此外,计算机可以嵌入在另一设备中,例如移动电话、个人数字助理(PDA)、移动音频或视频播放器、游戏控制台、全球定位系统(GPS)接收器或便携式存储设备,例如通用串行总线(USB)闪存驱动器,仅举几例。

[0121] 适用于存储计算机程序指令和数据的计算机可读介质包括所有形式的非易失性存储器、介质和存储器设备,包括例如半导体存储器设备,例如EPROM、EEPROM和闪存设备;磁盘,例如内部硬盘或可移动磁盘;磁光盘;以及CD ROM和DVD-ROM盘。

[0122] 为了提供与用户的交互,本说明书中描述的主题的实施例可以在具有用于向用户显示信息的显示设备(例如,CRT(阴极射线管)或LCD(液晶显示器)监视器)以及用户可以通过其向计算机提供输入的键盘和定点设备(例如,鼠标或轨迹球)的计算机上实现。其他种类的设备也可以用于提供与用户的交互;例如,提供给用户的反馈可以是任何形式的感觉反馈,例如视觉反馈、听觉反馈或触觉反馈;并且可以以任何形式接收来自用户的输入,包括声音、语音或触觉输入。另外,计算机可以通过向用户使用的设备发送文档和从用户使用的设备接收文档来与用户交互;例如,通过响应于从网络浏览器接收的请求将网页发送到用户设备上的网络浏览器。此外,计算机可以通过向个人设备(例如,运行消息传送应用的智能电话)发送文本消息或其他形式的消息并且作为回报从用户接收响应消息来与用户交互。

[0123] 用于实现机器学习模型的数据处理装置还可以包括例如专用硬件加速器单元,用于处理机器学习训练或生产的公共和计算密集部分,即推理、工作负载。

[0124] 可以使用机器学习框架(例如,TensorFlow框架、Microsoft Cognitive Toolkit框架、Apache Singa框架或Apache MXNet框架)来实现和部署机器学习模型。

[0125] 本说明书中描述的主题的实施例可以在计算系统中实现,该计算系统包括后端组件(例如,作为数据服务器),或者包括中间件组件(例如,应用服务器),或者包括前端组件(例如,具有图形用户界面、web浏览器或应用的客户端计算机,用户可以通过该图形用户界面、web浏览器或应用与本说明书中描述的主题的实施方式交互),或者一个或多个这样的后端、中间件或前端组件的任何组合。系统的组件可以通过任何形式或介质的数字数据通

信(例如,通信网络)互连。通信网络的示例包括局域网(LAN)和广域网(WAN),例如因特网。

[0126] 计算系统可以包括客户端和服务端。客户端和服务端通常彼此远离,并且通常通过通信网络进行交互。客户端和服务端的关系借助于在相应计算机上运行并且彼此具有客户端-服务端关系的计算机程序而产生。在一些实施例中,服务端将数据(例如,HTML页面)发送到用户设备,例如,用于向与充当客户端的设备交互的用户显示数据和从其接收用户输入的目的。可以在服务端处从设备接收在用户设备处生成的数据,例如,用户交互的结果。

[0127] 虽然本说明书包含许多具体的实施方式细节,但是这些不应被解释为对任何发明的范围或可能要求保护的范围的限制,而是作为可以特定于特定发明的特定实施例的特征的描述。在本说明书中在单独的实施例的上下文中描述的某些特征也可以在单个实施例中组合实现。相反,在单个实施例的上下文中描述的各种特征也可以单独地或以任何合适的子组合在多个实施例中实现。此外,尽管特征可以在上面被描述为以某些组合起作用并且甚至最初如此要求保护,但是来自所要求保护的组合的一个或多个特征在一些情况下可以从组合中删除,并且所要求保护的组合可以针对子组合或子组合的变型。

[0128] 类似地,虽然在附图中描绘了操作并且在权利要求中以特定顺序叙述了操作,但是这不应被理解为要求以所示的特定顺序或按顺序执行这些操作,或者执行所有示出的操作,以实现期望的结果。在某些情况下,多任务和并行处理可能是有利的。此外,上述实施例中的各种系统模块和组件的分离不应被理解为在所有实施例中都需要这种分离,并且应当理解,所描述的程序组件和系统通常可以一起集成在单个软件产品中或封装到多个软件产品中。

[0129] 已经描述了主题的特定实施例。其它实施例在所附权利要求书的范围内。例如,权利要求中记载的动作可以以不同的顺序执行并且仍然实现期望的结果。作为一个示例,附图中描绘的过程不一定需要所示的特定顺序或相继顺序来实现期望的结果。在一些情况下,多任务和并行处理可能是有利的。

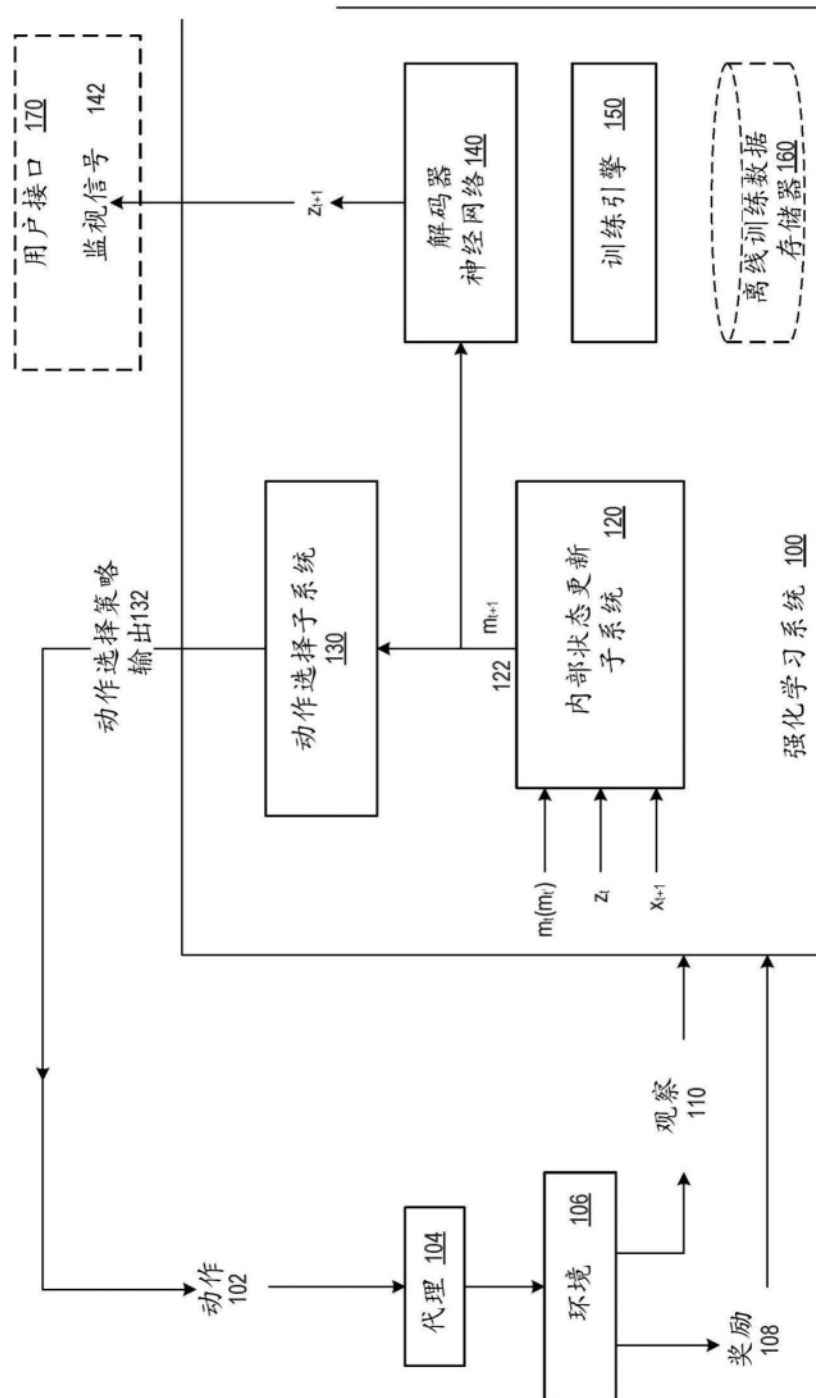


图1

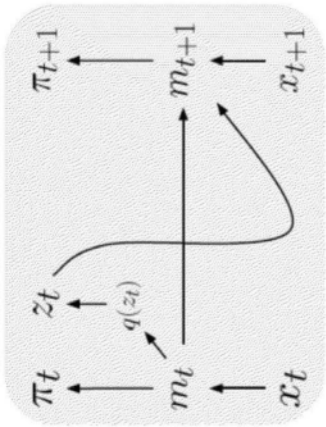


图2A

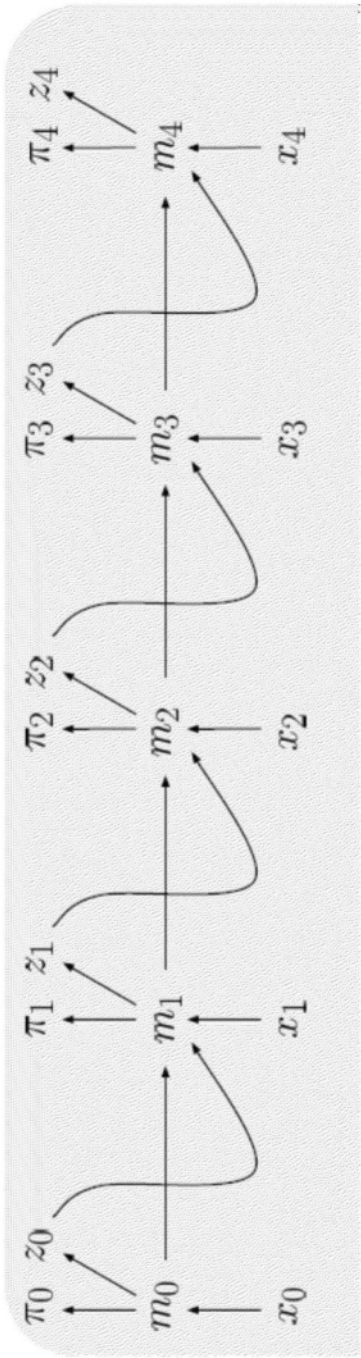


图2B

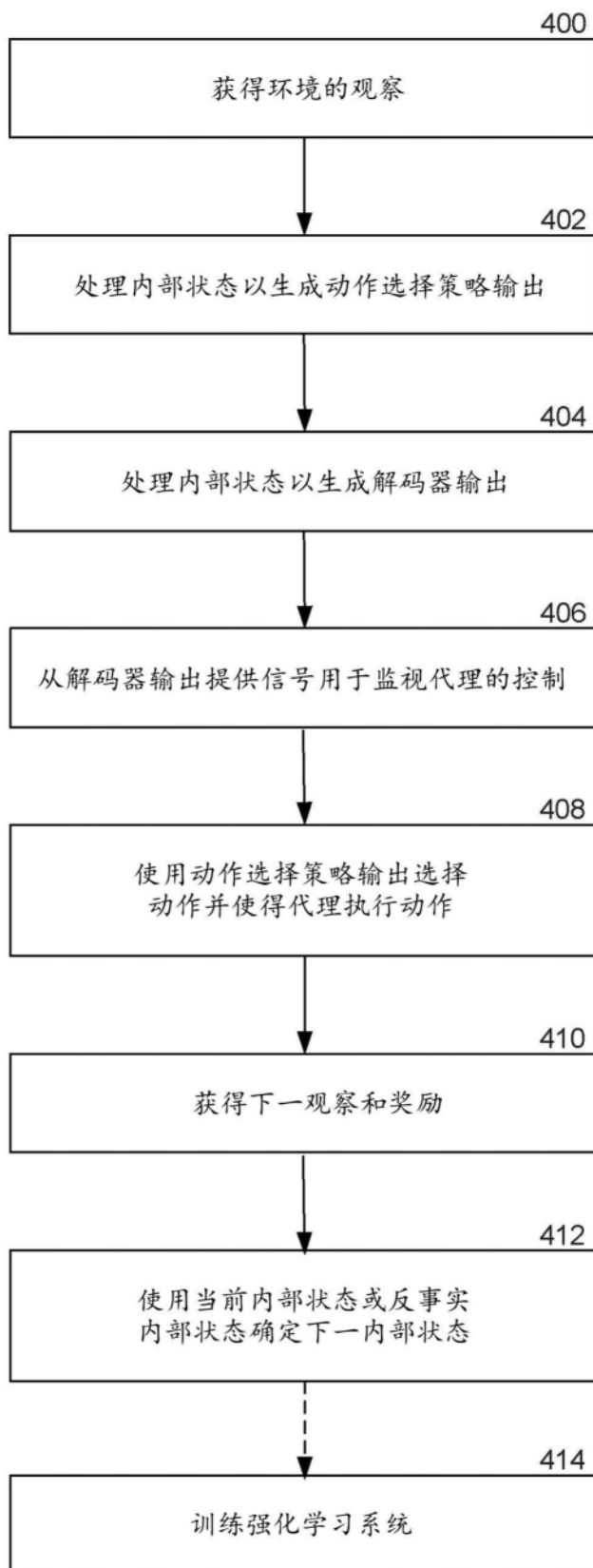


图4

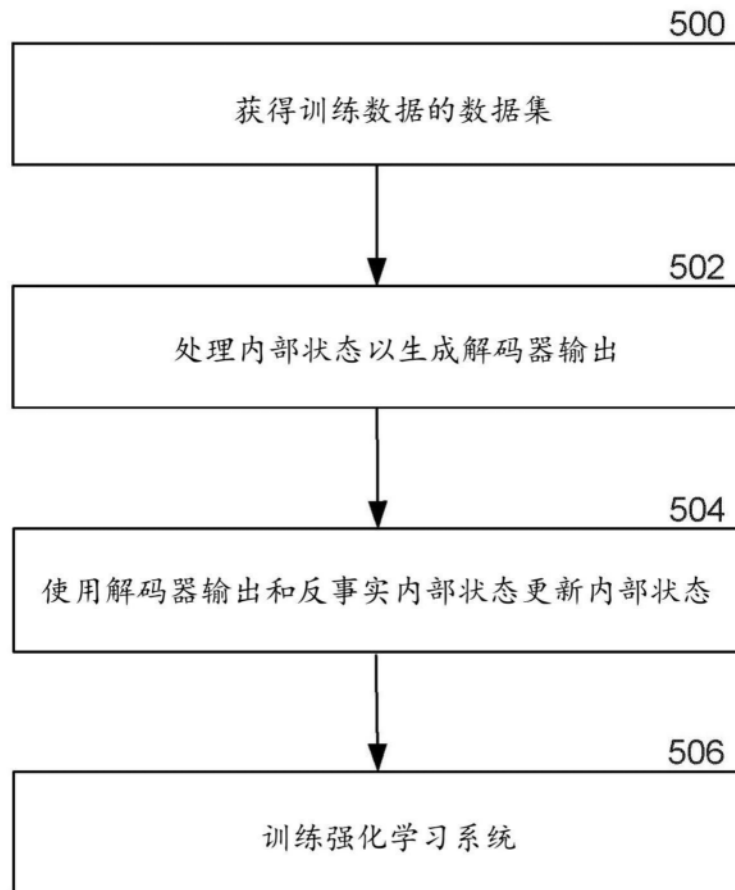


图5

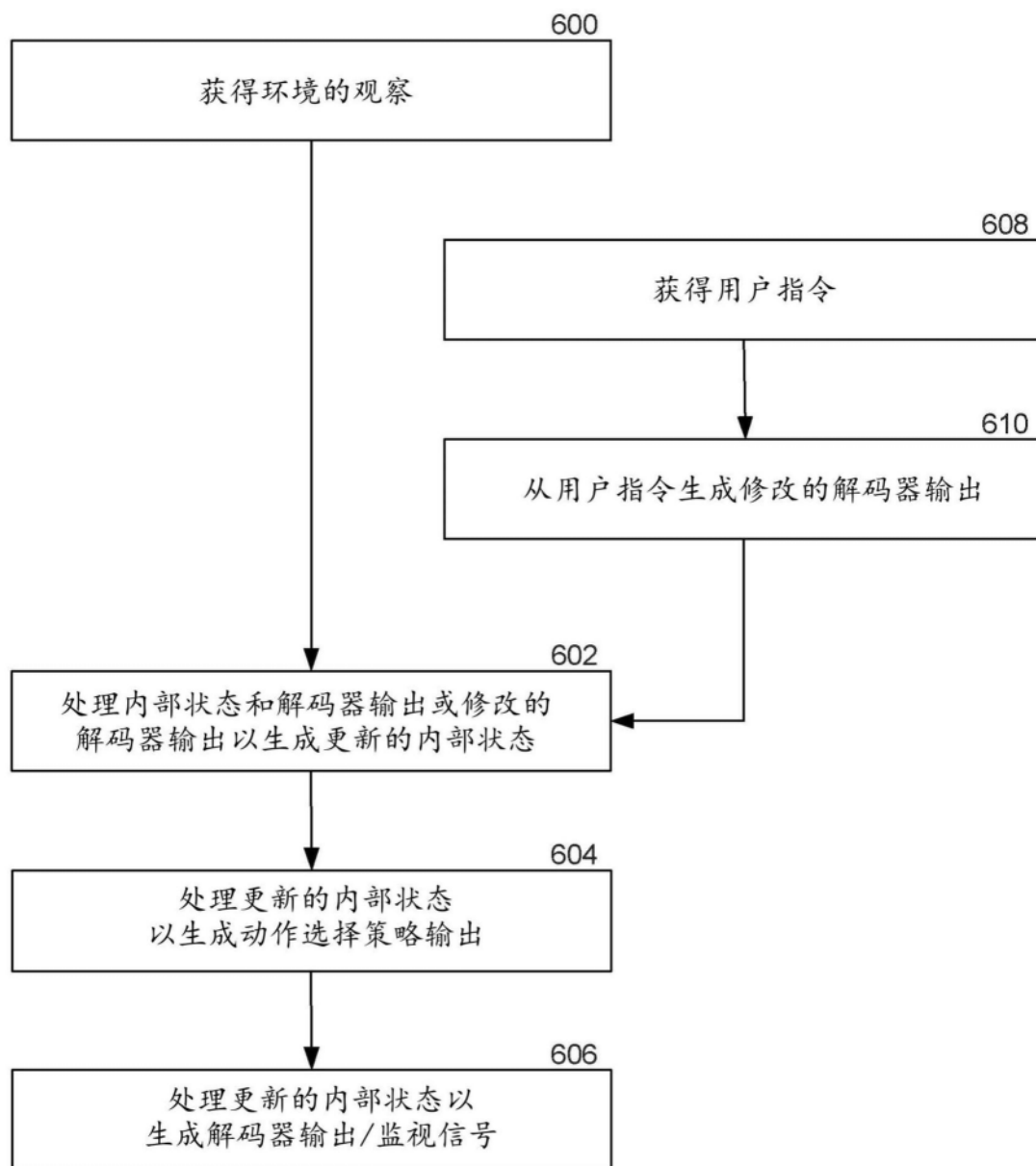


图6

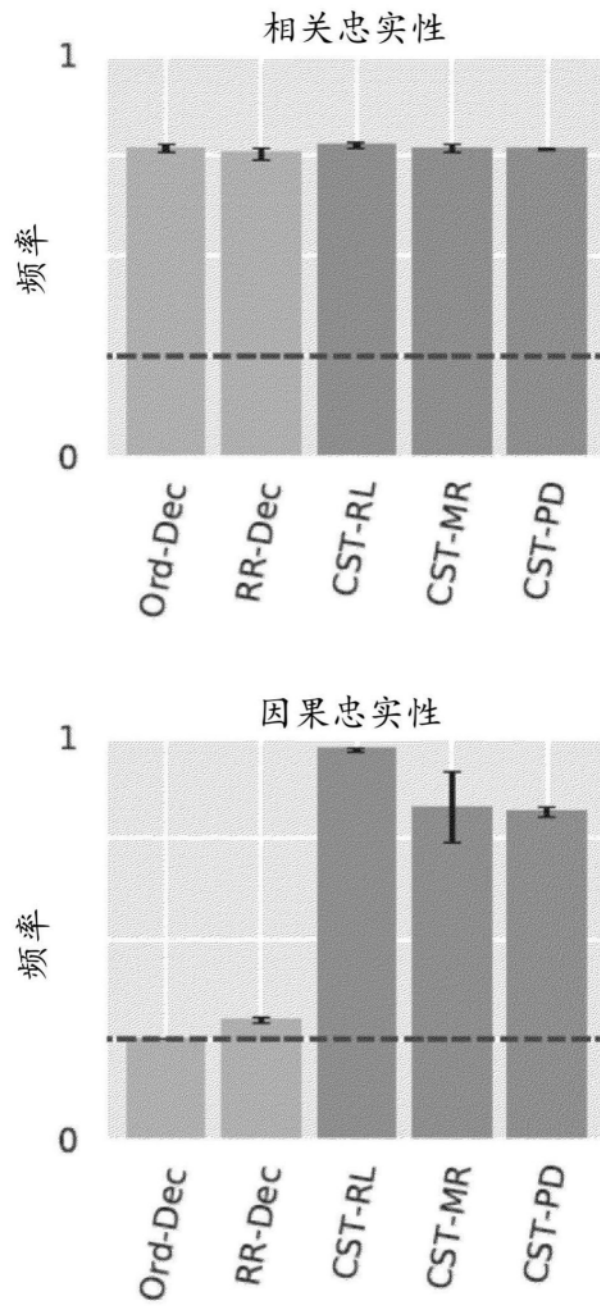


图7